FY 1996 Imputation and Estimation Plan

Quantum Research Corporation proposes to impute missing data for the FY 1996 R&D expenditures survey using the same techniques and algorithms that had been used by QRC for past sample surveys. These techniques involve:

- use of inflator/deflator factors to estimate key total variables, when necessary; and
- use of the previous year's questionnaire data as a reference distribution for allocating the key total variables among subtotal and detail questionnaire lines.

The inflator/deflator factors will be computed on the basis of unimputed matched data from “clean” institutions. Three sets of inflators/deflators will be computed, one for each of the following “key” variables:

Variable 1100-1	Total R&D Expenditures
Variable 1110-1	Total Federally Funded R&D Expenditures; and
Variable 1800-1	Total Current Fund Expenditures for Research Equipment.

The inflators for each of these variables will be separately computed for doctorate-granting institutions, master's-granting institutions, and other degree-granting institutions. These categories will be further subdivided into public and private institutions. The inflator/deflator factors will be applied to the previous year's key variables to obtain the current year estimates for nonrespondent institutions. For totally nonrespondent institutions, all three key variables will be estimated; for partially nonrespondent institutions only the missing key variables will be estimated.

This imputation technique is generally called ratio imputation and takes the following mathematical form:

$$\hat{y}_{ik_t} = \hat{B}_{k_t} \; y_{ik_{t-1}}$$

where $\hat{y}_{ik_t}$ is the estimated value of key variable y_k for institution *i* for year *t*,

$y_{ik_{t-1}}$ is the value of key variable y_k for institution *i* for year *t*-1, and

B_{k_i} is the inflator/deflator factor for key variable y_k , defined as

$$\hat{B}_{k_t} = \sum_{j=1}^{r} y_{jk_t} \Big/ \sum_{j=1}^{r} y_{jk_{t-1}}$$

where *r* is the set of institutions in the same degree level and institutional control peer group as institution *i* that provided key variable y_k in both years *t* and *t*-1.

In cases where the population of a class is less than 50 and too small to result in a meaningful inflator/deflator factor, the class will be combined with an adjacent class so that a reasonable inflator/deflator factor can be obtained. For the FY 1996 inflator/deflator factors, we propose to use the calculated values for public and private doctorate-granting institutions for all of the key variables. We expect to combine the data for public and private master's-granting institutions to calculate inflator/deflator factors for all of the key variables for master's-granting institutions. Likewise, we expect to combine the data from public and private other degree-granting institutions to calculate their key variables. It should be noted that the imputation rates for the key variables should be quite low again this year because of the high response rate we expect to achieve. Imputation rates for key variables in FY 1995 were 1.3 percent for total expenditures, 1.2 for Federal expenditures and 5.5 percent for current fund equipment expenditures.

The ratio imputation technique is used to impute key variables. Variables in the R&D expenditures survey, however, are hierarchical and each key variable has associated with it a considerable number of lower-level non-key variables. The key variable Federally-funded R&D expenditures, for example, has associated with it some 32 lower-level non-key variables such as Federally-funded R&D expenditures in astronomy, Federally-funded R&D expenditures in physics, and Federally-funded R&D expenditures in chemistry.

Non-key variables will be derived from their associated key variables using the relation:

$$\hat{y}_{ij_t} = \hat{y}_{ik_t}\left(\frac{y_{ij_{t-1}}}{y_{ik_{t-1}}}\right)$$

where $\hat{y}_{ij_t}$ is the estimated value of non-key variable y_j for institution *i* for year *t*,
$\hat{y}_{ik_t}$ is the estimated value of key variable y_k for institution *i* for year *t*,
$y_{ij_{t-1}}$ is the value of non-key variable y_j for institution *i* for year *t*-1, and
$y_{ik_{t-1}}$ is the value of key variable y_k for institution *i* for year *t*-1.

All imputed data cells will be given a status code of "I" to reflect their data source. If for a particular institution, lower-level non-key institutional data are not available for the previous year, all figures for that institution will be reviewed by the data collection coordinator and the data allocated based on data reported in earlier years or if necessary, data reported by peer institutions.

If an institution for which a given variable was previously imputed provides a response this year, the imputed value will be reestimated as follows:

$$\hat{\hat{y}}_{k_v} = y_{k_u} + \frac{v-u}{t-u}\left(y_{k_t} - y_{k_u}\right)$$

where $\hat{\hat{y}}_{k_v}$ is the reestimated value of imputed variable $\hat{y}_k$ for year *v*,
y_{k_u} is the reported value for variable y_k for earlier year *u*,
y_{k_t} is the reported value for variable y_k for current year *t*, and
$t > v > u$.

As far as our estimation plan is concerned, our estimation method is determined by the sample structure. The existing sample structure consists of five strata, four which are certainty strata and only one is a sampled stratum. All institutions in the sampled stratum population have an equal probability (.25) of being included in the sample. The estimation method used to estimate sampled stratum totals will thus be

$$\hat{Y}_{j_k} = \sum_{i=1}^{n_k}\left(y_{ij_k}\right)\left(\frac{N_k}{n_k}\right)$$

where $\hat{Y}_{j_k}$ is the estimated total for variable y_j in stratum *k*.

y_{ij_k} is the reported or imputed value for variable y_j for institution *i* in stratum *k*,

N_k is the total number of institutions in stratum *k*, and

n_k is the number of sampled institutions in stratum *k*.